TELL-Seq™ Data Analysis Software User Guide

for

Visualization on IGV

For Research Use Only. Not for use in diagnostic procedures.

Document # 100031 Version 2.0

March 2022

Table of Contents

1. Introduction

This document describes how to visualize TELL-Seq™ data on the Integrative Genomics Viewer (IGV).

The TELL-Seq™ WGS Library Prep kit uses an innovative Transposase Enzyme Linked Long-read Sequencing (TELL-Seq) technology to generate a paired-end library with barcoded linked reads suitable for Illumina® sequencing. Linked reads can be processed for variant calling, haplotype phasing, structural variation (SV) detection, metagenomic studies, and *de novo* sequencing assembly.

IGV is an open-source interactive tool developed by the Broad Institute for the visual exploration of genomic data [Ref.1].

This User Guide provides:

> **[1]** A description of the input files required for the generation of IGV tracks. The files are outputs from the Universal Sequencing Technology® (UST) Data Analysis Pipeline [Ref.2].
>
> **[2]** Directions on how to access and run the Perl script that generates IGV tracks (script name: *GenerateIGVFilesPseudoBamDelStyle.pl*).
>
> **[3]** Explanations on how to interpret IGV tracks displaying TELL-Seq data (including variant calls, haplotypes, and phase blocks).

This document does not describe how to run the UST Data Analysis Pipeline. This information can be found in the TELL-Seq Data Analysis Roadmap User Guide, TELL-Seq Data Analysis Software User Guide for Tell-Read, and TELL-Seq Data Analysis Software User Guide for Tell-Sort; the three guides available in the *User Guides* tab on the UST website [Ref.2]. This document does not describe either basic IGV usage. This information can be found on the IGV website [Ref.1].

2. Input Files

Three types of input files are required to generate IGV tracks using Tell-Sort-processed TELL-Seq data:

- Tell-Sort output **bam** files (one for each chromosome) in which duplicate reads have been removed and unique reads are ready for phasing analysis.
- Tell-Sort output **vcf** files (one for each chromosome) in which SNVs have been called and phased.
- The **reference genome** and its indexes.

3. Output Files

Four types of files are generated by the script that processes Tell-Sort outputs to generate IGV tracks:

3.1. Phased vcf and index files

This file contains Tell-Sort phased variants and their genomic location in the reference genome. Tell-Sort—software developed by UST to analyze TELL-Seq data [Ref.2]—generates vcf files for each chromosome. The *GenerateIGVFilesPseudoBamDelStyle.pl* script combines all vcf files into a single vcf file without changing their content, also creating an index file. The combined vcf file includes phased and unphased heterozygous positions with the "GT" attribute or genotype (assigned as "|" or "/" depending on whether the variant is phased or remains unphased, respectively). Each variant will also have the "PS" attribute for the phase set. The value will be the *phased block id* for phased variants; alternatively, there will be no value for unphased variants (assigned as ".").

3.2. Phased bam and index files

The phased-bam file contains Tell-Sort aligned reads with phasing information. The script combines the bam files for each chromosome to generate a single bam file containing all chromosomes. The phased information from the phased vcf file will be added to generate the phased-bam file. Compared to the unphased-bam file, the phased-bam file contains two customized tags: the HP tag, which assigns haplotypes ("1" or "2"); and the PS tag, which assigns phased sets (phased block id).

3.3. Gtf file

This file contains the imputed phase blocks aligned to the reference genome. The starting and ending coordinates of the block are obtained from the phase set in the phased vcf file. Phase blocks are shown as a horizontal bar on IGV spanning across the region with phased variants.

3.4. Pseudo-bam file

This file displays variants as part of haplotypes (phase blocks). This file translates the phased variants from the phased vcf file into pseudo reads (*contigs*) stored in bam format so that the user can upload them into the alignment track. The pseudo-bam file facilitates the visualization of the phased vcf file. In the pseudo-bam file, variants and genotypes are shown as vertical bars connected by a thick line and a number. The line connects variants that belong to the same haplotype (phase block). The number indicates the distance between adjacent phased positions. By default, reference alleles are shown in grey and alternate alleles are color-coded (set through Preferences, Alignments tab, and Show mismatched bases).

4. Script

1) Download script "GenerateIGVFilesPseudoBamDelStyle.pl" and example datasets from the Analysis Tools tab in the UST website under *Download Software for Visualization on IGV* (https://www.universalsequencing.com/downloadsoftware)
2) The code requires some dependencies. Users need to check whether these dependencies have been already installed and located in $PATH; or, alternatively, users should get and install these dependencies from:

a. Whatshap: https://whatshap.readthedocs.io/en/latest/installation.html
b. Samtools, bcftools, and htslib: http://www.htslib.org/download/

3) Running the script:

a. $ /path/to/GenerateIGVFilesPseudoBamDelStyle.pl -tellsortDataDir <TellsortDataDir> -prefix <prefixUsedInTellsort> -reference <referenceGenome> -maxCores <MaximumCores>
b. Example: $/data/yxia/bin/GenerateIGVFilesPseudoBamDelStyle.pl -tellsortDataDir . -prefix 501 -reference /data/genome/hg38/hg38.fasta -maxCores 30
    i. "/data/yxia/bin/GenerateIGVFilesPseudoBamDelStyle.pl": run this script.
    ii. "-tellsortDataDir .": the tellsort results are in the current directory, so "."
    iii. "-prefix 501": during the tellsort run, the "--prefix 501" has been used. Therefore, same thing needs to be provided to the current script.
    iv. "-reference /data/genome/hg38/hg38.fasta": during the tellsort run, the reference genome of "/data/genome/hg38/hg38.fasta" is configured with "--ref". Therefore, the same reference needs to be provided to the current script.
    v. "-maxCores 30": the maximum CPUs of 30 will be used in the process.

4) Result files: under the same directory of "-tellsortDataDir" provided, a sub-directory of prefix with "_IGV_Files" will be generated, the result files are saved here.

a. An example: with the command described above, a sub-directory of "501_IGV_Files" will be created, and under it:
    i. 501.gtf
    ii. 501.phased.bam
    iii. 501.phased.bam.bai
    iv. 501.phased.vcf.gz
    v. 501.phased.vcf.gz.tbi
    vi. 501.pseudo_bam.bam
    vii. 501.pseudo_bam.bam.bai
b. The meaning for each file in the example:
    i. "501.gtf": the gtf file for the phased block range on the reference genome.
    ii. "501.phased.bam" and "501.phased.bam.bai": the phased bam and its index.
    iii. "501.phased.vcf.gz" and "501.phased.vcf.gz.tbi": the phased vcf and its index.
    iv. "501.pseudo_bam.bam" and "501.pseudo_bam.bam.bai": the pseudo bam and its index.

5) Running performance: it will depend on the data size and the cores used. Using about 1 billion pair-end reads and 30 cores, the process takes approximately 2.5 hours.

5. IGV version

IGV version 2.4 or higher is required to view Tell-Sort data, as only these versions support to view the linked reads by parsing the bam tags of BX (barcode), MI (molecular identifier), and HP (haplotype).

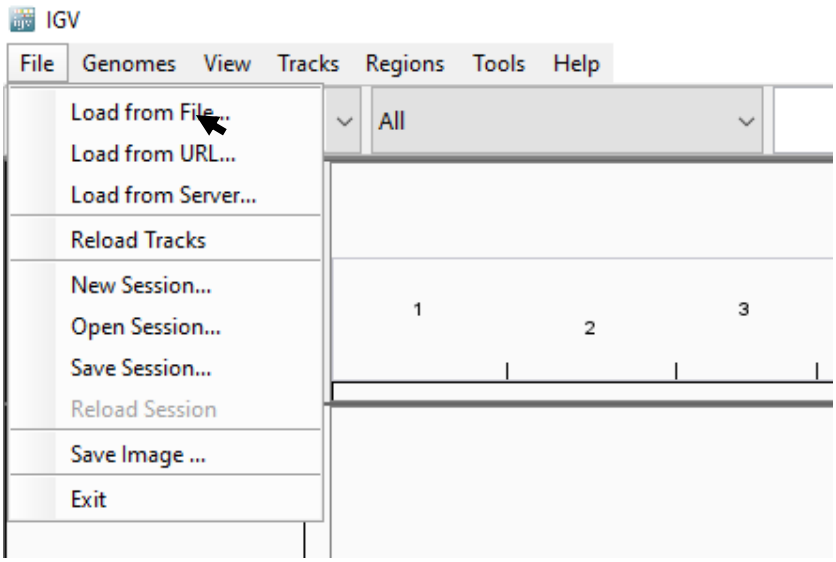
6. Viewing Tell-Sort data on IGV

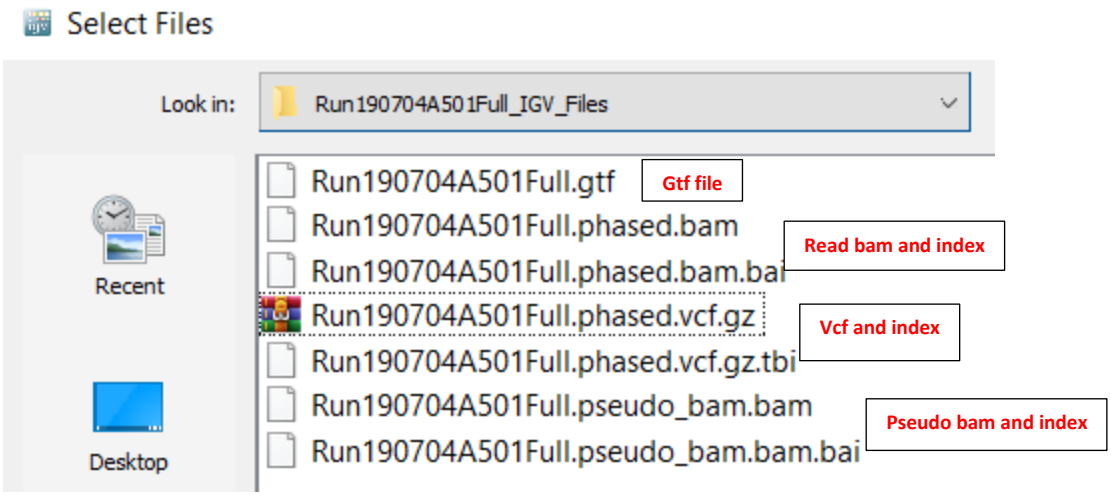
6.1. File loading

We recommend loading the four IGV tracks described here for a comprehensive visualization of TELL-Seq data (i.e., phased vcf, phased bam, gtf, and pseudo bam). However, when working with multiple samples, visualizing all tracks at the same time might not be possible. In this case, loading only the phased-bam and pseudo-bam files might be sufficient.

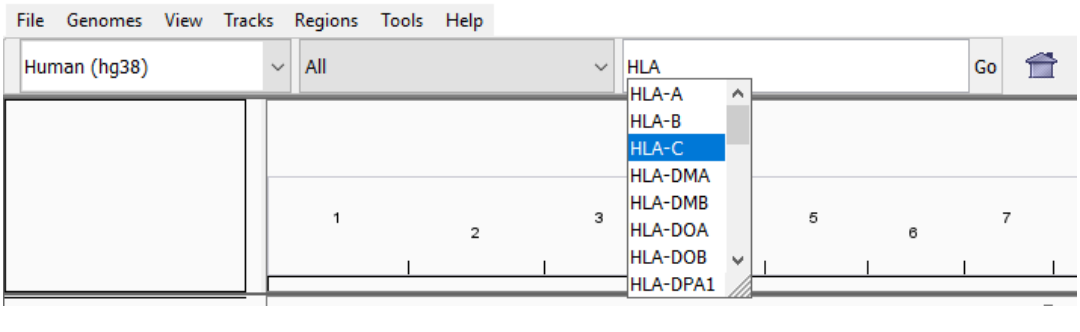
Before uploading the tracks, the user should make sure that the proper genome reference has been selected (in general, hg38). Alternatively, a different reference could be created by the user (for example, when phasing only the transcriptome/cDNA or when working with targeted applications).

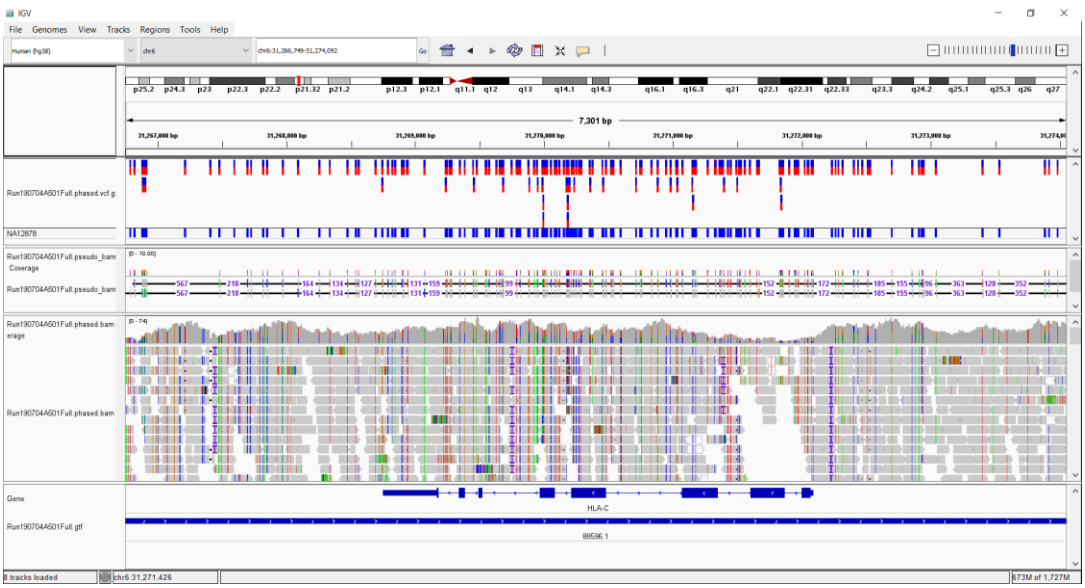
The following images represent the steps for uploading the four IGV tracks from a single sample:





The loading order and layout is flexible and will depend on the user preferences. We suggest the following layout (from top to bottom): vcf, pseudo bam, phased bam, and gtf files/tracks.
After uploading the files, the users should type the genomic region or the gene symbol of interest:

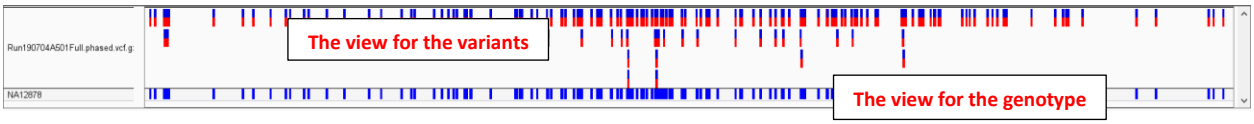




6.2. Understanding each track and adjusting the settings for optimum visualization

Vcf track

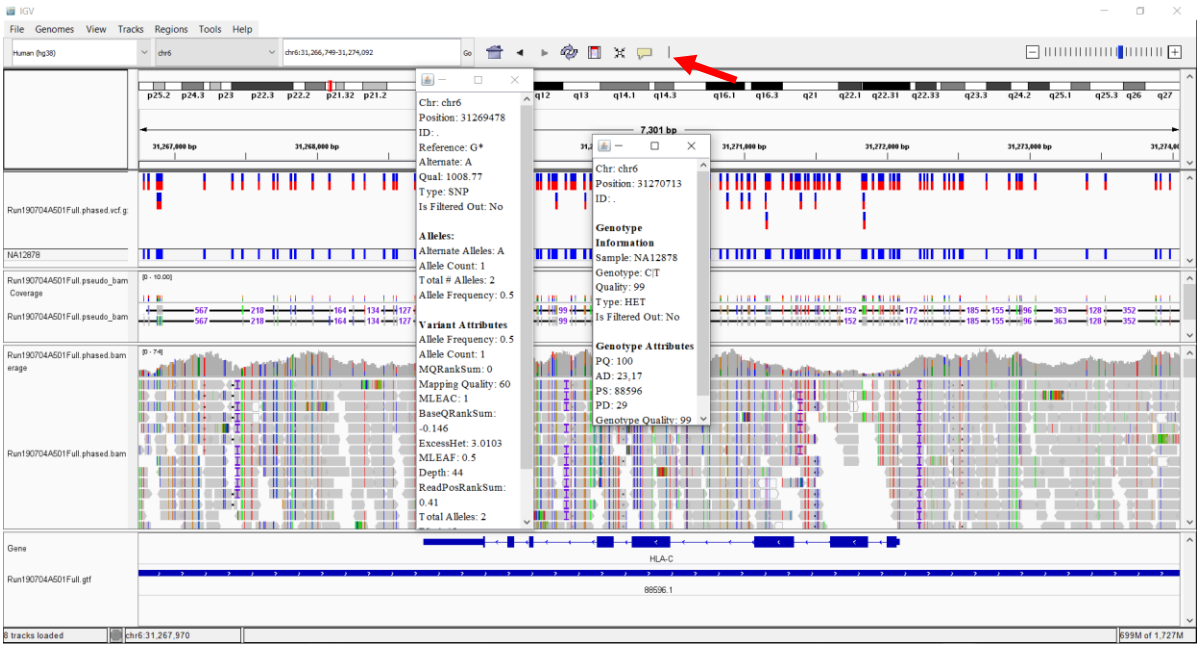
Tell-Sort provides a vcf file that can be uploaded as an IGV track. This track contains two track views: variants and genotypes. The users may need to re-size the track panel or scroll up and down to see all the views in this track:



The view for the variants shows vertical bars with color-coded calls relative to the reference genome at a given genomic position. Blue represents an allele identical to the reference. Red represents an alternate allele. In the current Tell-Sort setting, we have only included the heterozygous variants in the vcf file—shown as vertical blue and red bars. Occasionally, red-only bars may display. In these positions, the two alleles will be different from the reference (alternate). Depending on variant density, there might be multiple rows in this view to allow the users to easily find and interact the variant calls with the mouse.

The view for the genotype shows vertical blue bars in one row.

In the vcf track, both variant and genotype views will not show detailed information unless the user interacts the sites with the mouse (examples shown in the image below).
TIP: IGV allows users to modify pop-up text behavior in data panels by clicking where indicated by the red arrow.



Pseudo-bam track

Tell-Sort provides a pseudo-bam file that can be uploaded as an IGV track. As described above, this track converts variant genotype information into pseudo reads mapped to the reference. The pseudo-bam file includes two tags: “hp” for haplotype, and “PS” for the phase set. These tags display haplotypes in phase blocks.

After the pseudo-bam file has been uploaded on IGV, the view settings may need to be adjusted. For example, right click mouse on the track and select “Show all bases.” This action will allow IGV to display color-coded reference and alternate alleles.

In this alignment track, since pseudo reads derive from the genotypes in the phased vcf file and each read represents an allele in the giving phased block, the view for the coverage on the top of the track will not be much meaningful. The value of this track is to provide users with a convenient way to read the phased genotypes. Since genotypes are color-coded, users will be able to easily detect phased variants in the same phase block.

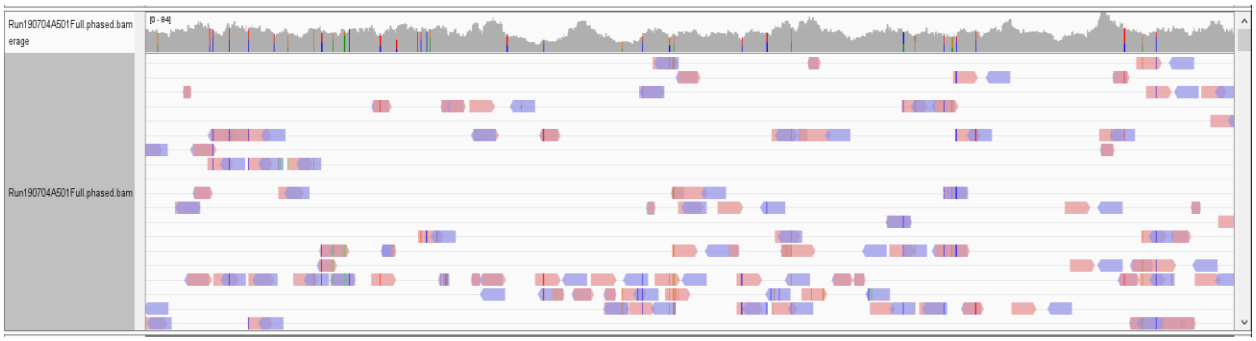
Read bam track

Tell-Sort provides a read bam file that can be uploaded as an IGV track to show how reads have been mapped to the reference. The read bam file includes three tags for linked reads: the BX tag for the barcode; the HP tag for the haplotype; and the PS tag for the phase set. (Note: The BX tag is added to the bam file when linked reads are aligned to the reference, while the phasing information tags HP and PS are added to the bam file when running open-source WhatsHap for haplotag [Ref.3]). This track

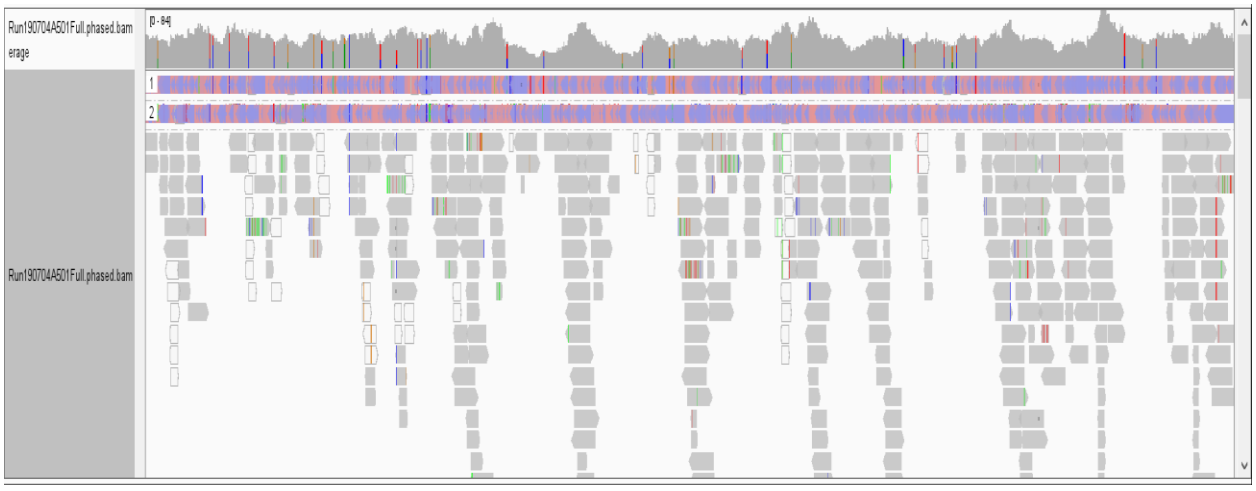
shows how linked reads have been mapped; which reads have been assigned to haplotypes 1 and 2, or remain unphased; and which reads belong to a phase block.

The track has two views: coverage (on top), which is not different to shot-gun sequences; and read alignments (at the bottom). Depending on the purpose, the view settings may need to be adjusted as it follows:

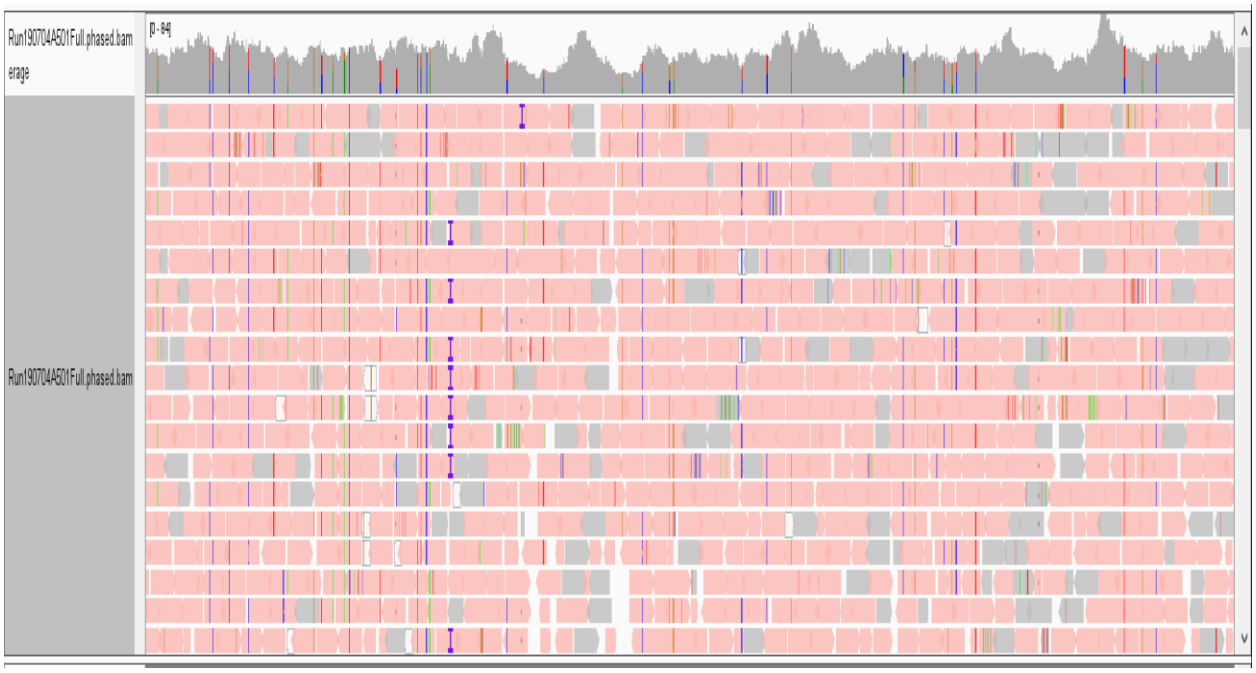
- **View the reads linked by barcodes**: right click in the track, select “Link by BX” or “Link by tag…”, and type “BX” in the text box; then click “OK.”



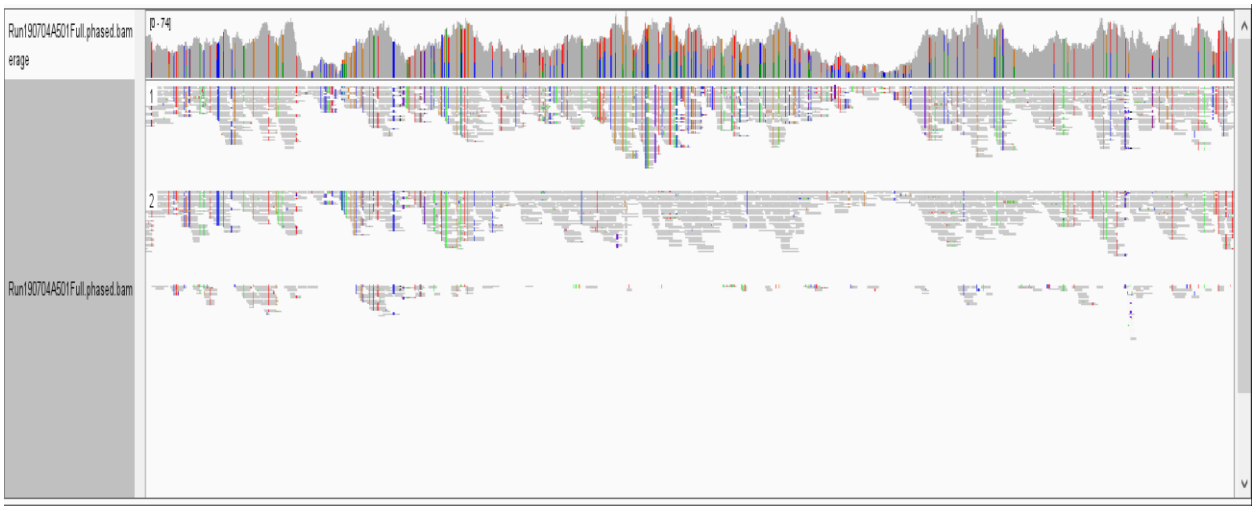
- **View the reads linked by haplotype**: right click in the track, select “Link by HP” or “Link by tag…”, and type “HP” in the text box; then click “OK”. Right click, select “Group alignments by”, and select “phase.”



- **View reads belonging to the same phased block by colors**: right click in the track, select “Color alignments by,” then select “tag,” and type “PS” in the text box; click “OK.”



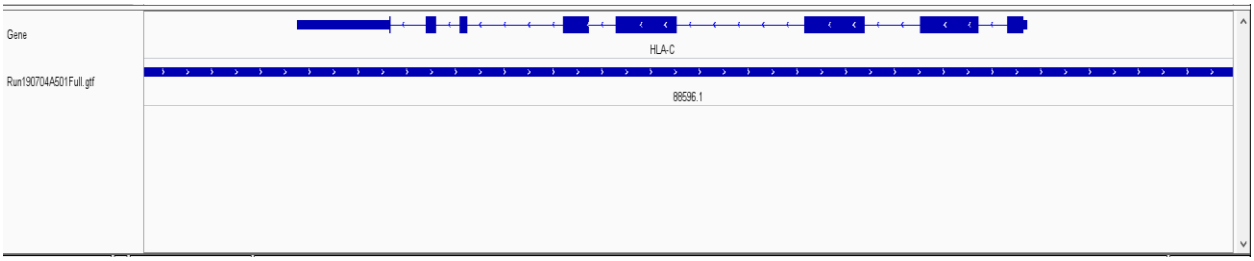
- **Linked read view with the reads in each haplotype as well as in the unphased**: right click in the track, and select “Linked read view (MI).” Right click, select “Group alignments by,” and select “phase” (or select “tag”, and in the text box, type “HP”, and click “OK”).



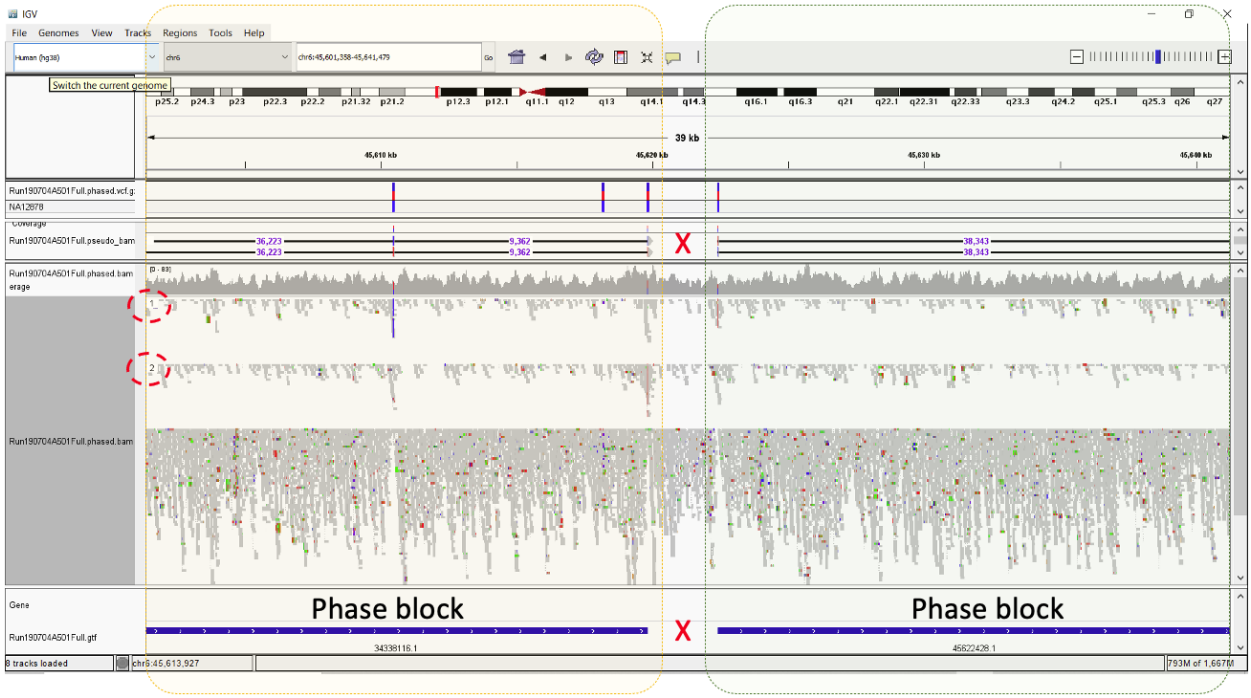
- There are additional view setting configurations in IGV. Users may need to customize the settings to fit their own specific purposes.

Gtf track

Tell-Sort generates the gtf file, which displays phase blocks on IGV as horizontal bars underneath a subtract showing gene annotations. See an example in the image below.



TIP: Users should be aware that IGV-assigned haplotypes (either haplotype 1 or 2) cannot be interpreted beyond the limits of every individual phase block. For example, the image below shows two phased blocks (see the pseudo-bam track or the gtf track). Haplotypes 1 and 2 are assigned in the phased-bam track (red circles). However, haplotype 1 in the left phase block may or may not be present in the same chromosomal copy (maternal or paternal) than haplotype 1 in the right phase block, as there are not phased variants connecting both blocks (see "X").



7. References

1) Integrative Genomics Viewer (IGV) https://software.broadinstitute.org/software/igv/
2) TELL-Seq Software User Guides https://www.universalsequencing.com/protocol-gate
3) WhatsHap tool https://whatshap.readthedocs.io/en/latest/

Revision History

Document #	Version	DCR Reference and Comment
100031-USG	1.0	DCR-210100 Initial Release
100031-USG	1.1	Adding instructions to upload script